Huang *et al.*
EURASIP Journal on Audio, Speech, and Music Processing (2023) 2023:45
https://doi.org/10.1186/s13636-023-00312-8

EURASIP Journal on Audio, Speech, and Music Processing

EMPIRICAL RESEARCH

Open Access

W2VC: WavLM representation based one-shot voice conversion with gradient reversal distillation and CTC supervision

Hao Huang[1,2], Lin Wang[1], Jichen Yang[3*], Ying Hu[1] and Liang He[1,4]

Abstract

Non-parallel data voice conversion (VC) has achieved considerable breakthroughs due to self-supervised pre-trained representation (SSPR) being used in recent years. Features extracted by the pre-trained model are expected to contain more content information. However, in common VC with SSPR, there is no special implementation to remove speaker information in the content representation extraction by SSPR, which prevents further purification of the speaker information from SSPR representation. Moreover, in conventional VC, Mel-spectrogram is often selected as the reconstructed acoustic feature, which is not consistent with the input of the content encoder and results in some information lost. Motivated by the above, we proposed W2VC to settle the issues. W2VC consists of three parts: (1) We reconstruct feature from WavLM representation (WLMR) that is more consistent with the input of content encoder; (2) Connectionist temporal classification (CTC) is used to align content representation and text context from phoneme level, content encoder plus gradient reversal layer (GRL) based speaker classifier are used to remove speaker information in the content representation extraction; (3) WLMR-based HiFi-GAN is trained to convert WLMR to waveform speech. VC experimental results show that GRL can purify well the content information of the self-supervised model. The GRL purification and CTC supervision on the content encoder are complementary in improving the VC performance. Moreover, the synthesized speech using the WLMR retrained vocoder achieves better results in both subjective and objective evaluation. The proposed method is evaluated on the VCTK and CMU databases. It is shown the method achieves 8.901 in objective MCD, 4.45 in speech naturalness, and 3.62 in speaker similarity of subjective MOS score, which is superior to the baseline.

Keywords Voice conversion, Self-supervised pre-trained Representation, Gradient reversal layer (GRL), CTC

1 Introduction

Voice conversion aims to convert the speech of a source speaker into a simulation of the speech of a target speaker, while retaining the linguistic information unchanged. VC has been used in many applications, such as speaker-identity modification for text-to-speech [1], singing voice conversion [2], and generation of various kinds of expressive speech [3].

According to the existence of parallel utterance pairs from both the source speaker and target speaker in the training dataset, VC can be broadly divided into parallel VC and non-parallel VC. In recent years, non-parallel VC has become the mainstream research topic because no

*Correspondence:
Jichen Yang
NisonYoung@163.com
[1] School of Computer Science and Technology, Xinjiang University, Urumqi, China
[2] Xinjiang Key Laboratory of Multi-lingual Information Technology, Urumqi, China
[3] School of Cyber Security, Guangdong Polytechnic Normal University, Guangzhou, China
[4] Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing, China

parallel training data is necessary and thus can be easily implemented.

In non-parallel VC research, variational autoencoder (VAE) [4, 5] has become increasingly popular, which commonly employs an encoder-decoder framework to learn the disentanglement of the content representation and speaker representation from source speaker and target speaker respectively. It can match the task of VC that disentangles the content information of source speaker and speaker information of target speaker first, then combine them, and finally synthesize the converted utterance.

In conventional VAE-based VC, the encoder extracts speaker-independent latent variables from the input acoustic features, and the decoder reconstructs the acoustic features from the latent variables and the given speaker representations. However, the modeling of VAE is limited to Gaussian distributional assumption [6], and the distribution of actual data may be much more complicated. On the basis of VAE, any-to-any (also called one-shot VC) VC was achieved earlier in vector quantized variational autoencoder (VQ-VAE) [7] and AUTOVC [8]. They both relied on the disentanglement of content and speaker information within an utterance with a content encoder and a speaker encoder. VQ-VAE used learning the prior via an embedding dictionary to represent entirely different phonetic contents, while AUTOVC used a pre-trained speaker encoder [9] to obtain speaker information and used an information bottleneck to limit the leakage of the source speaker information. In addition, phoneme posteriorgram (PPG) is very popular in VAE-based VC [10]. The reason is that PPG is speaker-independent content representation in theory. In recent VAE-based VC studies, self-supervised pre-trained representation have been used to extract content or speaker information because SSPR is often trained from a large-scale unlabeled data and hence SSPR contains abundant and compact speech information [11]. SSPR have been used to provide content information for VC tasks [12]. Several SSPRs have been used as the inputs of content encoders in [12]. For example, FragmentVC applies the wav2vec 2.0 as the input of content encoders [13]. In S2VC [14], CPC [15] is used as the input of both the content encoder and the speaker encoder.

In the early VAE-VC, the traditional Mel-spectrogram is often selected as the input acoustic features and reconstruction features, which are fed into the encoder and obtained from the decoder, respectively. Then various vocoders such as WaveNet [16], MelGAN [17], and HiFi-GAN vocoder [18] are used to convert Mel-spectrogram to waveform speech. Even in recent VC studies using SSPR [11, 13, 14] as the input feature of content representation extraction for source speaker, Mel-spectrogram is also selected as the reconstructed acoustic feature converted from transformed SSPR. Though SSPRs have been used in previous work of voice conversion such as [11, 13, 14] and made promising progress in VC performance, there are still two limitations in current system configurations:

- There is no special tactic to remove speaker information in content representation extraction for the source speaker when SSPR is used as the input. As a result, there is still more or less speaker information in the content representation.
- Mel-spectrogram is often selected as the reconstructed representation by decoding the combining content representation and speaker representation (denoted transformed-SSPR hereafter), there will inevitably be some information missing in the process of converting transformed-SSPR to Mel-spectrogram.

In order to settle the problems in VC based on SSPR mentioned above, several strategies are proposed.

(1) Connectionist temporal classification (CTC) [19] is used to align content representation and text context from the phoneme level, and content encoder plus gradient reversal layer (GRL)-based speaker classifier [20, 21] are used to remove speaker information from content representation with WavLM representation (WLMR) [22] as the input. The reason is that the GRL-based speaker classifier can make the content information not flow to the speaker classifier, which makes the output of the content encoder have no speaker information. Finally, WLMR can be used as the input to extract content representation because it can make great success in speech recognition [22].
(2) WLMR is selected as the reconstructed representation rather than the commonly used Mel-spectrogram, which can be consistent with the input of the content representation to escape the information lost.
(3) WLMR-based HiFi-GAN vocoder is trained to convert the reconstructed WLMR to waveform speech.

In this regard, a VC method based on WLMR is proposed in this work, which is termed as W2VC for short. The contribution of the work can be described as follows:

- Content encoder plus GRL-based speaker classifier are used to remove speaker information in the content representation extraction of source speaker.

- The reconstructed feature is WLMR rather than commonly acoustic features such as Mel-spectrogram. We compare the quality of the reconstructed Mel-spectrogram synthesized audio obtained by the traditional VAE structure, using this novel WLMR that can escape the information lost in the process of converting transformed-SSPR to acoustic feature.
- WLMR-based HiFi-GAN vocoder is trained to convert synthesized WLMR to waveform speech. We retrain the HiFi-GAN vocoder using WLMR instead of Mel-spectrogram obtained from speech signal processing, resulting in a direct mapping from sampling points to audio. The experimental results show that the quality of the audio generated by the retrained vocoder is greatly improved. The problem of inconsistent input features between WLMR and conventional vocoder is solved.

The rest of the paper is organized as follows: Section 2 introduces the proposed method. Section 3 gives the evaluation and analysis. Section 4 draws the conclusion.

2 Methods

We first described the proposed W2VC in detail. Figure 1 demonstrates the framework. As shown, the proposed W2VC includes a content encoder, a speaker encoder, a decoder, and a retrain vocoder. In which, we adopt the pre-trained self-supervised model WavLM to learn the representation and feed the resulting WLMR to the content encoder to extract the content representation. The speaker embedding is obtained from the speaker encoder. The CTC auxiliary module and the GRL-based speaker classifier are used to remove speaker information in the content representation. It is worth noting that these two kinds of auxiliary networks are only used in the training stage and not in the inference stage. The two representations are concatenated into the decoder to obtain the reconstructed representation and then sent to the retraining vocoder to synthesize speech. Next, they will be introduced in detail one by one.

2.1 WavLM representations

In this paper, we employ the well-known pre-trained model WavLM to learn the representation. This new WLMR replaces the traditional acoustic features as input to the system. The schematic diagram of WavLM is shown in Fig. 2. The pre-trained WavLM model uses a masked speech denoising and prediction framework, where some of the inputs are simulated noise/overlapped with the mask, and the goal is to predict pseudo-labels of the original speech in the masked region. The WavLM model uses a pre-training scheme for speech modeling with denoising masks. As shown below, the WavLM model consists of a convolutional encoder and a Transformer encoder. Among them, the convolutional encoder has seven layers, and each layer contains a time domain

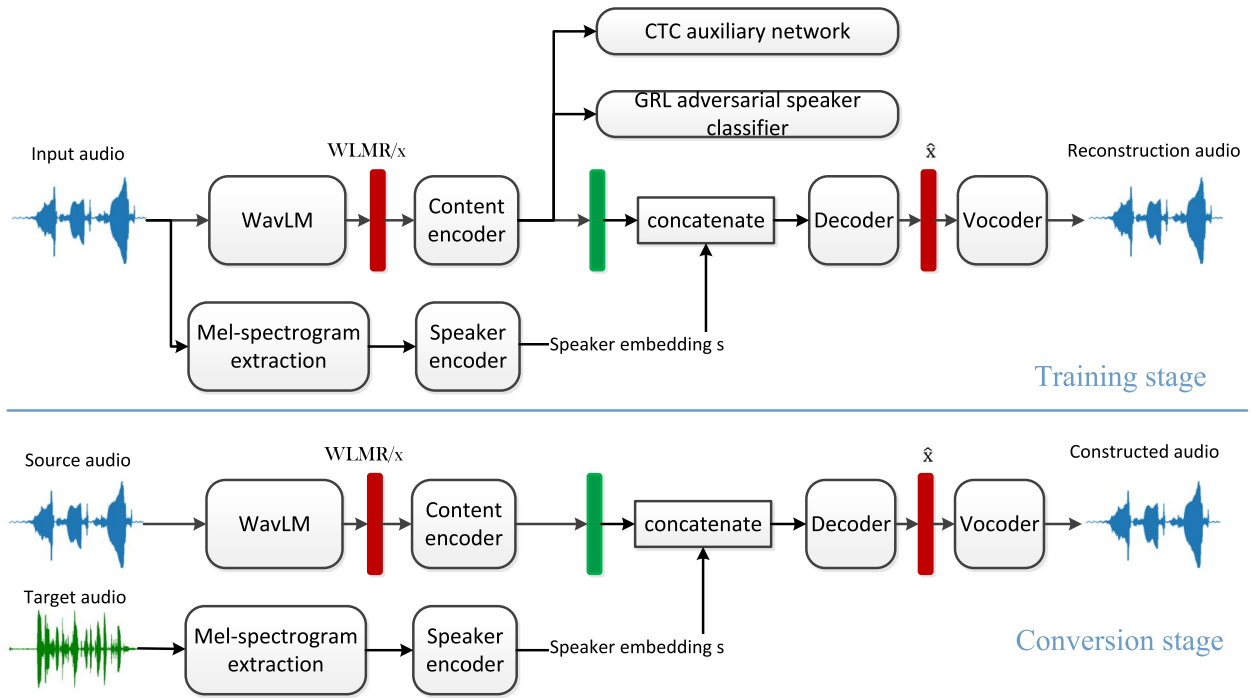

Fig. 1 The framework of W2VC

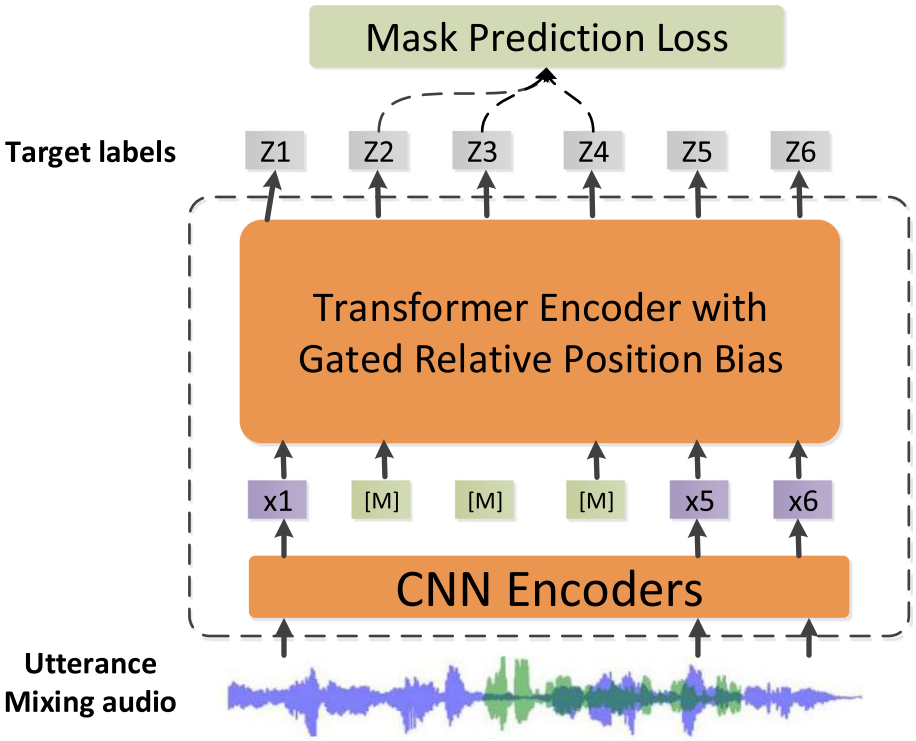

Fig. 2 The architecture of the WavLM model

convolutional layer, a normalization layer, and a GELU activation function layer. The Transformer encoder has 24 encoder layers, 1024-dimensional hidden states, and 12 attention heads, resulting in 316.62M parameters. The relative position embedding is shared by all layers, so it does not significantly increase the number of parameters. Researchers use gated relative position bias to introduce relative position into the calculation of attention networks, so as to better model local information. During training, WavLM randomly transforms the input waves, for example mixing two waves, or adding background noise. Then, about 50% of the audio signal is randomly masked, and the label corresponding to the mask position is predicted at output. WavLM draws on the idea of HuBERT [23] and converts the continuous signal into discrete markers by the K-means method and models the discrete markers as targets.

2.2 Content represesentation extraction

Content representation extraction is used to extract content representation from WLMR and it is realized by three modules which consist of the content encoder, CTC auxiliary networks and GRL-based speaker classifier. This part describes the network structure of the content encoder and its two auxiliary networks.

2.2.1 Content encoder

The content encoder is shown in Fig. 3. According to [24, 25], it can be known that the effectiveness of instance normalization for disentanglement in computer vision has been verified. Instance normalization in [26] works well to separate speaker and content representations in one-shot voice conversion. We improve the content encoder in [26] with four convolutional blocks with one more block and a bottleneck linear layer. The instance normalization layer is used in the content encoder to normalize the global information and effectively extract the content representation. The content encoder takes the WLMR as input from which the content information is decoupled and used as a preliminary extraction of the content embedding.

2.2.2 CTC auxlilary network

CTC is a loss function of neural networks, which is often used to solve the problem of time series mapping with uncertain alignment between input features and output labels. More specifically, the CTC loss function is modeled by an auxiliary network that enables the content encoder to learn content representation that contains only textual information. The CTC-aided network design aims to improve the decoupling ability of the content

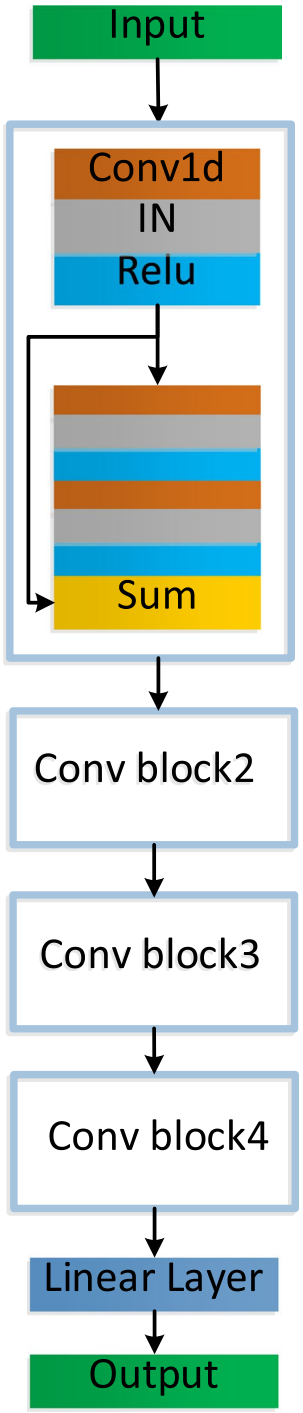

Fig. 3 Block diagram of the content encoder

encoder from the content information by using stronger textual supervision.

Supposing the feature obtained from WavLM pre-trained model can be denoted as x. Then x is fed into the content encoder to obtain $z(x)$. Given an $z(x)$ input sequence $I = [i_1, i_2, ..., i_T]$, where T is the length of the input sequence. The task of the CTC loss function is to maximize the probability distribution $P(m|I)$ for the corresponding phoneme label sequence m of length. Here, the length we use is the CMU dictionary plus a blank label from CTC is 40. CTC represents $P(m|I)$ as a summation of all possible frame-level intermediate representations $\omega = [\omega_1, \omega_2, ..., \omega_T]$. The CTC loss is defined as:

$$L_{CTC} = P_{max}(m|I) = \sum_{\omega \in \phi(m)} P(\omega|I) \tag{1}$$

2.2.3 GRL-based speaker classifier

GRL-based speaker classifier is used to remove speaker information in the content representation. Inspired by [27], an adversarial prediction network [28] is designed to reduce the information overlap between the speaker embedding and the content embedding. The common classifier is to pass the loss layer by layer in the back propagation, and then each layer of the network will calculate the gradient according to the returned error, and then update the parameters of the network layer to realize the classification task. What the GRL-based speaker classification network does is to multiply the error passed to this layer by a negative number λ, which makes the network before and after GRL have opposite training objectives to achieve an adversarial effect. By adding GRL loss to guide the content encoder, the embedding learned contains more speaker-independent information. The GRL-based speaker classifier loss is defined as follows:

$$L_{GRL} = L_{grl-cls}(\theta_e, \theta_{S+G}) = -\sum_{k=1}^{K} II(u==k) \log p_k, \tag{2}$$

where $II(\cdot)$ is the indicator function, K is the number of speakers and u denotes the speaker who produced speech x, p_k is the probability of speaker k, θ_e denotes the parameters of the content encoder, and θ_{S+G} denotes the parameters of the GRL-based speaker classifier.

2.3 Speaker encoder

The goal of the speaker encoder is to generate the same embedding for different utterances from the same speaker, but different embeddings for different speakers. Conventionally, one-hot embedding of speaker identity is used in many-to-many VC. To build zero-shot VC, the speaker encoder needs to produce an embedding that is generalizable to unseen speakers. In this work we use Global Style Tokens (GST) [29] which has also been widely used speech synthesis [30, 31], as the speaker encoder. GST has the ability to compress a variable-length acoustic feature sequence into a fixed-dimensional

speaker reference embedding and learn the speaker embedding without an explicit style label. As shown in Fig. 4, two steps are needed for GST to extract speaker reference embedding:

(1) The reference encoder compresses the acoustic feature sequence by using continuously stacked Conv2d layers. It then takes the last state of a bidirectional GRU (Bi-GRU) layer as the output of the reference encoder. The variable-length acoustic feature sequence is therefore transformed into a fixed-length embedding, which is denoted as the reference embedding.
(2) The reference embedding is passed to a style token layer. In the style token layer, several equal-length style tokens (for example, A, B, C, D) are initialized to simulate different components of speech (for example, emotion, rhythm, pitch, and speech rate). Then, a multi-head self-attention module is constructed, which learns a similarity measure between the reference embedding and each style token. The attention module outputs a set of combination weights that represent the contribution of each style token to the reference embedding. The weighted sum of the style tokens is named as the reference embedding.

2.4 Decoder

The decoder takes the outputs of all encoders as input and reconstructs the representation, and feeds the decoded representation into the reconstructed vocoder. In terms of implementation details, the network architecture used in the experiments is shown in Fig. 5. The decoder contains 3 Conv1d layers, a BLSTM layer, and a linear layer. Each Conv1d layer has a kernel size of 5, a channel number of 512 and a stride of 1. The number of units of the BLSTM layer is 512.

2.5 Vocoder

The reconstructed feature in this work is WLMR, which has more comprehensive information than that in Mel-spectrogram and also is also consistent with the input of content encoder. In order to convert the reconstructed WLMR to waveform speech, we trained a HiFi-GAN vocoder with WLMR as input and waveform speech as output here.

2.6 Loss function

The total loss function has three parts: the first is used to calculate the loss between the input WLMR and the reconstructed WLMR, the second is used to calculate the loss of CTC, the third is GLR-based speaker classification, which can be written as:

$$L_{VC} = L_{MSE}(x, \hat{x}) + \alpha L_{CTC} + \beta L_{GRL}, \tag{3}$$

where x and $\hat{x}$ stand for the input WLMR and the reconstructed WLMR respectively, L_{MSE} is the mean-square-error (MSE) between x and $\hat{x}$, α, and β are hyper-parameters of L_{CTC} and L_{GRL}, respectively.

3 Evaluation and analysis

In this section, the proposed method is evaluated and corresponding analysis is given. Next, the used database, experimental setup, evaluation and analysis will be introduced one by one.

3.1 Database

The VCTK [32] and the CMU ARCTIC [33] corpora were used to evaluate the proposed method. There are 109 English speakers with different accents in the VCTK database and 43,398 utterances were obtained as the training set after pre-processing. Different from VCTK, CMU ARCTIC is a parallel English corpus. We performed both intra-gender and inter-gender voice conversion experiments and tested four transformations for each model :rms→slt(male to female), rms→bdl(male to male), clb→slt(female to female), and clb→bdl(female to male). All four sets of speakers were invisible during training, and 40 sentences were randomly selected for each speaker in the test set. Twenty volunteers were recruited for each subjective assessment test. For a

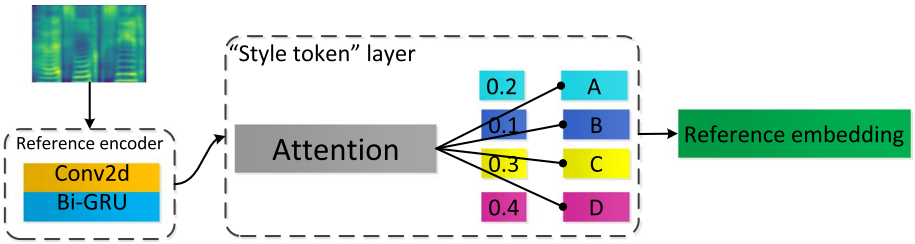

Fig. 4 Block diagram of speaker encoder. The input features are fed to the reference encoder followed by a style token layer to get the reference embedding

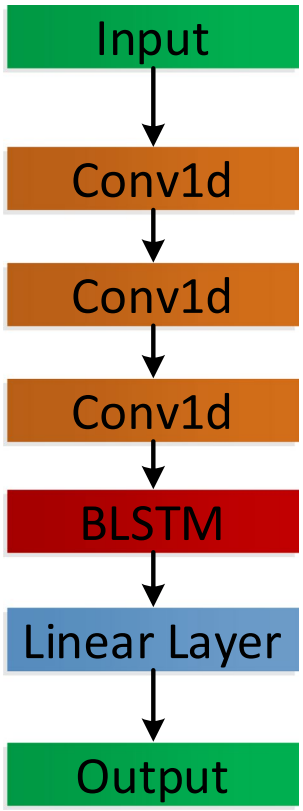

Fig. 5 Block diagram of the decoder

random sample from the CMU dataset, participants were asked to rate the naturalness of speech on a scale from 1 to 5. Higher scores represent higher speech naturalness.

The male speaker "bdl" and the female speaker "slt" were selected as the target speakers. The male speaker "rms" and the female speaker "clb" were selected as the source speakers. In addition, both VCTK and CMU ARCTIC were used for training and test. In addition, we selected a portion of the CMU Arctic corpus to evaluate our approach, including 40 utterances per speaker. Firstly, the waveform data is selected from the corpus and converted to 16 kHz. Then, a WavLM Large pre-trained model extracts a 1024-dimensional representation and sends it to the content encoder, while an 80 dimensional Mel-spectrogram is extracted and sent to the speaker encoder.

3.2 Experimental setup

3.2.1 Model configurations

(1) *Speaker and content encoder*: Our architecture uses the same speaker encoder as [29], which ultimately produces 256-dimensional speaker embeddings. The content representation is generated by 4 convolution blocks and a bottleneck linear layer. Each convolution block contains 3 Conv1d layers with skip connection, and each Conv1d layer combines with the instance normalization and the ReLU activation function. The kernel size, channel, and stride of the Conv1d layer are set to 5, 256, and 1, respectively. The resulting 1,024-dimensional representation is fed to the decoder. It's worth noting that the CTC auxiliary network and adversarial learning networks are only used during the training phase. The CTC auxiliary network is shown in Fig. 6, which contains 3 Conv1d layers, a BLSTM layer, and a linear layer. Each Conv1d layer is combined with batch normalization and a ReLU activation

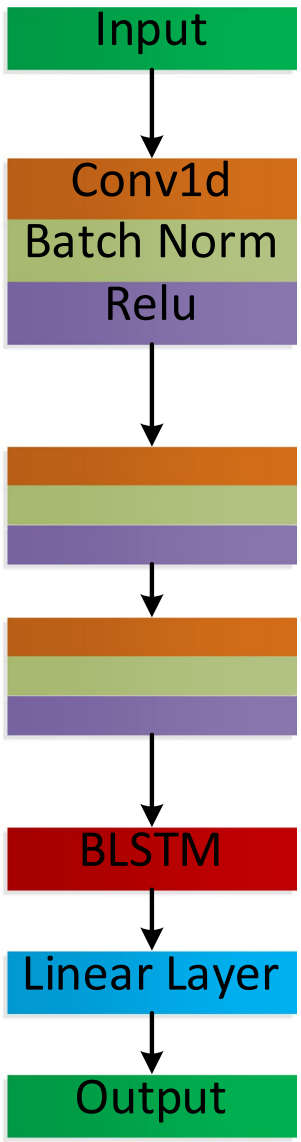

Fig. 6 Block diagram of the CTC model

layer. The number of BLSTM layer and linear layer units are set to 128 and 40, respectively. The 40 outputs of the linear layer correspond to the 39 English phonemes and an extra blank phoneme. Adversarial classifier is shown in Fig. 7, which consists of 2 activation functions, 2 linear layers for normalization, and a common speaker classifier. In the backword, gradient inversion is used.

(2) *WLMR*: In the experiments, we choose the WavLM Large models pre-trained for 1M and 700k steps on 94k large-scale diverse data using the labels generated by clustering the 9th transformer layer output of the released 2nd-iteration HuBERT Base model to get the 1024-dimensional bottleneck features.

(3) *Decoder*: The decoder is used to reconstruct the WLMR, which is composed of 3 Conv1d layers, a BLSTM layer, and a linear layer. Each Conv1d layer has a kernel size of 5, a channel number of 512, and a stride of 1. The number of units of the BLSTM layer is 512. The final linear layer maps the output of the BLSTM layer to reconstructed features.

(4) *Retrained HiFi-GAN*: In recent years, GAN-based vocoders have been more and more applied to speech synthesis, such as MelGAN, which can achieve CPU real-time synthesis. However, MelGAN still needs to be improved in sound quality and synthesis speed. Compared with the traditional GAN-based MelGAN vocoder, HiFi-GAN adds a Multi-Period Discriminator (MPD) on the basis of retaining the MelGAN multi-scale discriminator Multi-Scale Discriminator (MSD) and applies a multi-receptive field fusion module in the generator. Speech quality and reasoning speed are further improved. The advantage of the HiFi-GAN vocoder is that it can compress the audio signal while maintaining a high-quality audio output, which is useful for applications that require high-fidelity audio signals. In our experiments, we retrained the HiFi-GAN vocoder with the aim of allowing the representation extracted by the pre-trained model combined with the speaker features to better form the audio. This makes the vocoder change from Mel-spectrogram to audio synthesis in the traditional sense to adoption point to audio synthesis. We trained a multi-speaker HiFi-GAN neural vocoder using the features extracted by the WavLM Large pre-trained model on the VCTK corpus. The generated essential features are converted into waveform. The segment size of HiFi-GAN is set to 8192 and the sampling rate is 16k.

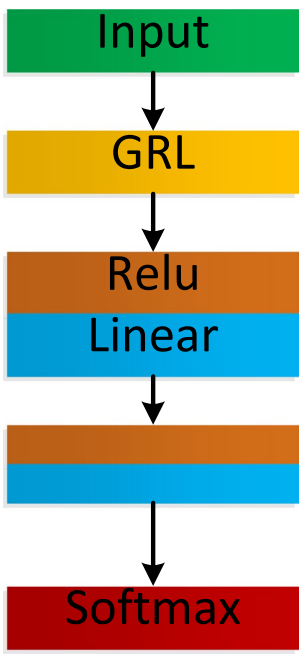

Fig. 7 Block diagram of GRL-based speaker classifier

To fit our corpus, we change the upsampling rate and kernel size of the first layer in the generator network to 10 and 20, respectively, in order to adapt to the feature length extracted by WavLM can be upsampled to the length of wav. Other model configuration methods are the same as [18].

ADAM optimizer with a learning rate of 0.0003 is used to train the model, batch size was set to 16. In the process of model training, the loss weight of L_{CTC} in Equation (3) increases linearly with the increase of training steps, and it will remain fixed after arrival, with a maximum value of 0.001 within 50k steps. The loss weight α of L_{GRL} is set to 1 and the number of training steps is set to 300K.

3.2.2 Training details

The following are the three baselines that are to be used for comparisons:

(1) *VQ-VAE baseline*: We learn from the model in paper [8]. The system is based on a non-parallel corpus, the acoustic feature is the Mel-spectrum feature, and the model structure includes four parts: content encoder, embedding dictionary, decoder, and GST timbre encoder. We train the VQ-VAE baseline model with a learning rate of 0.0003 using an ADAM optimizer. The Mini-batch size is set to 8. The number of VQ embeddings is set to 40 and the number of training steps is set to 300K.

(2) *CTC-VQ-VAE baseline*: We reproduce the model of [34], which consists of a content encoder with CTC

auxiliary loss, a vector quantizer, a speaker encoder, and a decoder. The weight of CTC loss increases linearly during model training, the maximum value within 50k steps is 0.001.

(3) *FragmentVC*:The SOTA approach FragmentVC [13] is considered as a baseline here. FragmentVC is proposed to obtain the latent phonetic structure of the source speaker's speech using the pre-trained model and the timbre features of the target speaker using log Mel-spectrograms. The alignment of two different feature Spaces is realized by adding an attention mechanism. We use the WavLM pre-trained model instead of the wav2vec2.0 [35] model, and the remaining configurations are consistent with [13].

When analyzing the results of the voice conversion system, the evaluation criteria are divided into two types: objective evaluation and subjective evaluation. Objective evaluation refers to the application of some measurement methods according to the results of the experiment, while subjective evaluation refers to the evaluation of the naturalness and similarity of the speech according to the subjective consciousness of the evaluator.

3.3 Subjective evaluation

To evaluate the overall speech synthesis quality we collect Mean Opinion Score (MOS) for all the neural models outputs as well as ground truth recordings. Speech quality includes naturalness and similarity. We performed both intra-gender and inter-gender voice conversion experiments, and for each model we tested four transitions: rms→slt(male to female), rms→bdl(male to male), clb→slt(female to female), and clb→bdl(female to male). All four groups of speakers were invisible during training, and 40 utterances were randomly selected for each speaker in the test set. We recruited 20 volunteers for each subjective evaluation test. For a random sample from CMU dataset, participants were asked to rate the naturalness of the speech on a 5 point scale from 1 to 5.

The MOS results with 95% confidence intervals of speech naturalness and speaker similarity are shown in Tables 1 and 2, respectively. The above two tables count the subjective evaluation results of speech naturalness and similarity of four models: VQ-VAE, CTC-VQ-VAE, FragmentVC, and W2VC in three aspects: intra-gender, inter-gender, and average. The proposed method, W2VC, significantly outperforms the other methods on both intra-gender and inter-gender voice conversion and achieves very good scores, with an average naturalness score (4.45 ± 0.1970) and an average speaker similarity score (3.62 ± 0.2153). CTC-VQ-VAE is superior to the VQ-VAE model in speech naturalness and speaker similarity, and the CTC text supervision information added by CTC-VQ-VAE is beneficial to content modeling. In addition, compared with the other three models, the poor performance of FragmentVC in speech naturalness may be caused by the fact that the model uses a retrained HiFi-GAN vocoder, and the features obtained by the decoder do not match the input features of the vocoder.

Table 1 The MOS results with 95% confidence intervals show the impact of VQ-VAE, CTC-VQ-VAE, FragmentVC, and the proposed method on speech naturalness

Method	Intra-gender	Inter-gender	Average
VQ-VAE	2.08 ± 0.1912	2.18 ± 0.2297	2.13 ± 0.2817
CTC-VQ-VAE	3.36 ± 0.1546	3.22 ± 0.1754	3.29 ± 0.1465
FragmentVC	1.54 ± 0.1497	1.54 ± 0.2216	1.54 ± 0.2437
W2VC	4.42 ± 0.1737	4.48 ± 0.1632	4.45 ± 0.1970

Table 2 The MOS results with 95% confidence intervals showing the impact of VQ-VAE, CTC-VQ-VAE, FragmentVC, and the proposed method on speech similarity

Method	Intra-gender	Inter-gender	Average
VQ-VAE	1.97 ± 0.2736	1.91 ± 0.2576	1.94 ± 0.2113
CTC-VQ-VAE	3.22 ± 0.2897	3.23 ± 0.2812	3.22 ± 0.2476
FragmentVC	2.34 ± 0.3666	2.42 ± 0.3878	2.38 ± 0.2635
W2VC	3.46 ± 0.2571	3.78 ± 0.2859	3.62 ± 0.2153

3.4 Objective evaluation

Mel cepstral distortion (MCD) is used to evaluate the naturalness for objective evaluation, wherein, lower MCD indicates better naturalness. In each set of VC systems, we used four pairs of transformations (clb-slt, clb-bdl, rms-slt, rms-bdl) and generated 40 utterances for each pair of transformations, totaling 160 utterances. We validate the results of MCD with 160 utterances produced by each VC system.

The objective evaluation results are shown in Table 3. From the results in the table, it can be seen that the MCD value of W2VC is lower than that of VQ-VAE and CTC-VQ-VAE. In other words, the performance of using CTC-assisted network and GRL-based speaker classification network is better than that of using CTC alone and introducing embedding dictionary or only introducing embedding dictionary. Because adding these two auxiliary networks enables the content encoder to learn more content information. This table also shows that the mapping between content and audio using the attention mechanism is the worst.

Table 3 Objective evaluation results comparison among VQ-VAE, CTC-VQ-VAE, FragmentVC, and the proposed method in terms of MCD

Method	MCD(dB)
VQ-VAE	9.716
CTC-VQ-VAE	8.988
FragmentVC	10.204
W2VC	8.901

3.5 Comparison with MelGAN vocoder

As mentioned above, WLMR-based HiFi-GAN vocoder is used to convert the synthesized WLMR to waveform in W2VC. Here we compare the performace between traditional vocoder such as MelGAN and the WLMR-based HiFi-GAN vocoder. To this end, we show in Table 4 the comparison of naturalness, similarity and MCD between speech formed using WLMR-based HiFi-GAN and the MelGAN vocoder. Note that, in order to use MelGAN vocoder, the synthesised feature by using the decoder is Mel-spectrogram rather than WLMR.

As shown in Table 4, the synthesized speech by the HiFi-GAN vocoder is much more better than that generated by MelGAN in naturalness and similarity. The same conclusion also can be found from the objective evaluation. We attribute this because some information is more or less lost in the process of Mel-spectrogram generation because the input feature is WLMR.

3.6 Speaker classification

To verify that the speaker encoder can effectively extract speaker information, we conducted speaker classification experiments on the speaker embeddings extracted by the speaker encoder. The speaker embedding is generated from the selected sentence, which represents the information about the speaker extracted from the sentence. The ability of a speaker classifier to classify speaker embeddings was used as a measure of speaker modeling ability. The desired result is that different utterances from the same speaker can be clustered together, and different speakers can be separated from each other. Better clustering indicates better results of speaker classification experiments, thus proving that the speaker encoder can effectively extract speaker information. The speaker classifier consists of three Conv1d layers, a BLSTM layer, and a linear layer. Each convolutional layer is paired with batch normalization and Relu activation functions, and the number of BLSTM nodes is 128. The activation function of the linear layer is softmax and the loss function is cross-entropy.

Figure 8 shows the vector comparison of CTC-VQ-VAE and the proposed W2VC with t-SNE in 2D space, where (a) and (b) denote CTC-VQ-VAE and W2VC, respectively. The same color represents the same speaker. For the speaker classification experiment, we selected five female (p225, p228, p229, p230, p231) and five male (p226, p227, p232, p237, p241) speakers from the VCTK corpus, each with 10 utterances for classification. The selected utterance is passed through the GST timbre encoder to generate its timbre vector. As can be seen from Figure (a) and Figure (b), the 10 utterances of the same speaker are well clustered and can be clearly separated between different speakers, which indicates that the speaker encoders of both systems have learned meaningful speaker embeddings and the utterances of different speakers are well distinguished from each other. It indicates that the embedding extracted by GST is able to serve as speaker identity. In addition, compared to Figure (a), the sentences of the same speaker are more closely distributed in Figure (b). Therefore, compared with CTC-VQ-VAE, W2VC has better clustering performance in terms of different speaker identities.

3.7 Visual analysis

We carried out a visual analysis of the W2VC model before and after conversion. The ground truth and the converted speech utterance obtained by W2VC were visualized in the Mel-spectrogram. Figure 9 shows a Mel-spectrogram comparison of the W2VC model for same-gender and cross-gender conversions. We use the source speaker "rms" for providing converted speech. Figure 9a demonstrates the Mel-spectrogram of the ground truth (arctic_a0002.wav) of the source speaker "rms". Figure 9b illustrates the results of cross-gender male-to-female conversion ("rms"-"slt"). Figure 9c shows the result of the same-gender male-to-male conversion ("rms"-"bdl").

Table 4 The MOS results with 95% confidence intervals in naturalness, similarity and MCD of the generated speech using WLMR-based HiFi-GAN and that using the MelGAN vocoder

Vocoder	Nat.			Sim.			MCD
	Intra	Inter	Avg	Intra	Inter	Avg	
MelGAN	2.27 ± 0.2364	1.98 ± 0.3334	2.13 ± 0.2692	2.45 ± 0.3133	1.83 ± 0.4296	2.14 ± 0.2589	9.508
HiFi-GAN	4.42 ± 0.1737	4.48 ± 0.1632	4.45 ± 0.1970	3.46 ± 0.2571	3.78 ± 0.2859	3.62 ± 0.2153	8.901

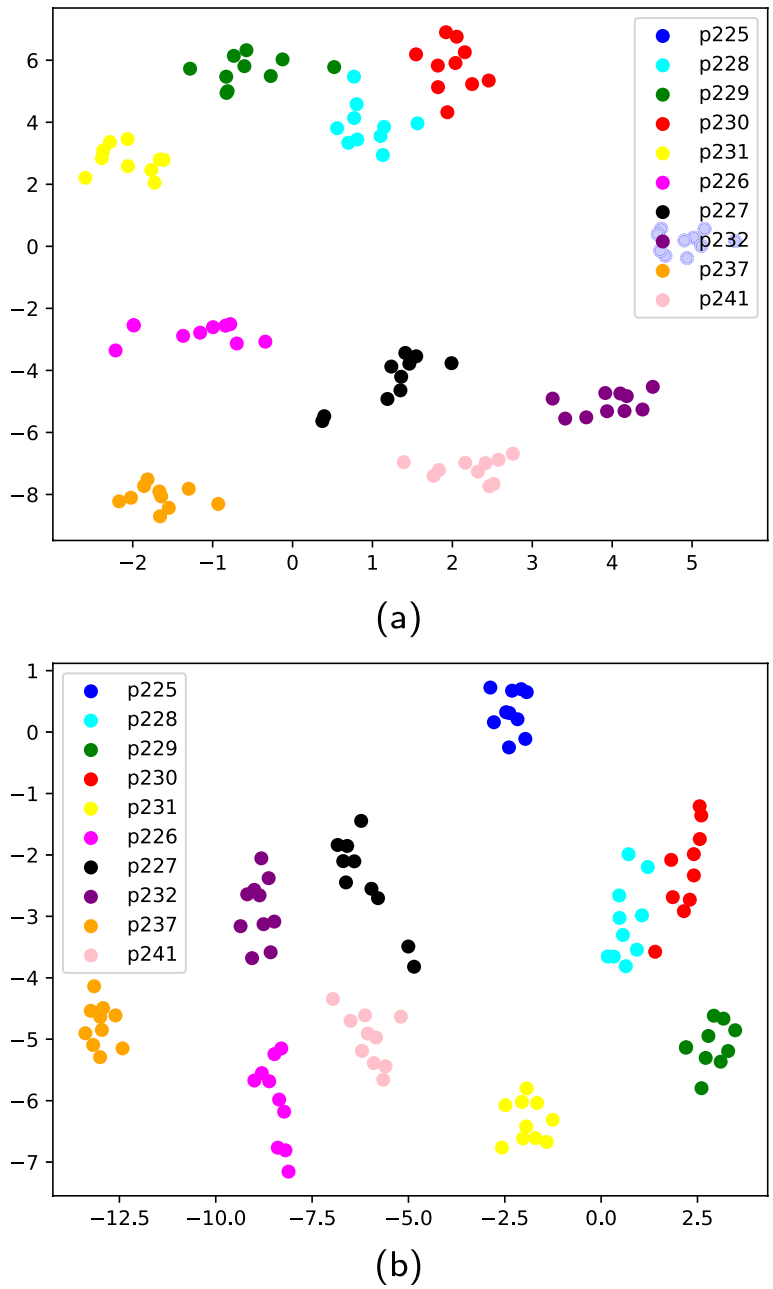

Fig. 8 Visualization of speaker embedding comparison between CTC-VQ-VAE (**a**) and W2VC (**b**)

As can be seen in Fig. 9, the Mel-spectrogram outline of the converted speech using W2VC is consistent with that of the ground truth, which indicates that the converted speech content is preserved. Comparing Fig. 9b and c with a, respectively, the brighter points in the figure represent higher decibels. The comparison between (b) and (a) at the same frequency shows that the identity information such as timbre and pitch converted to the "slt" of the target speaker is also guaranteed. By comparing (a) with (c), we can see that the brightness and pitch of male-to-male ("rms"-"bdl") are basically the same, indicating better performance in male-to-male conversion.

3.8 Ablation study

We perform ablation studies on the GRL-based speaker classifier network and the CTC-assisted network based on W2VC. We conducted a subjective evaluation of MOS scores including naturalness and speaker

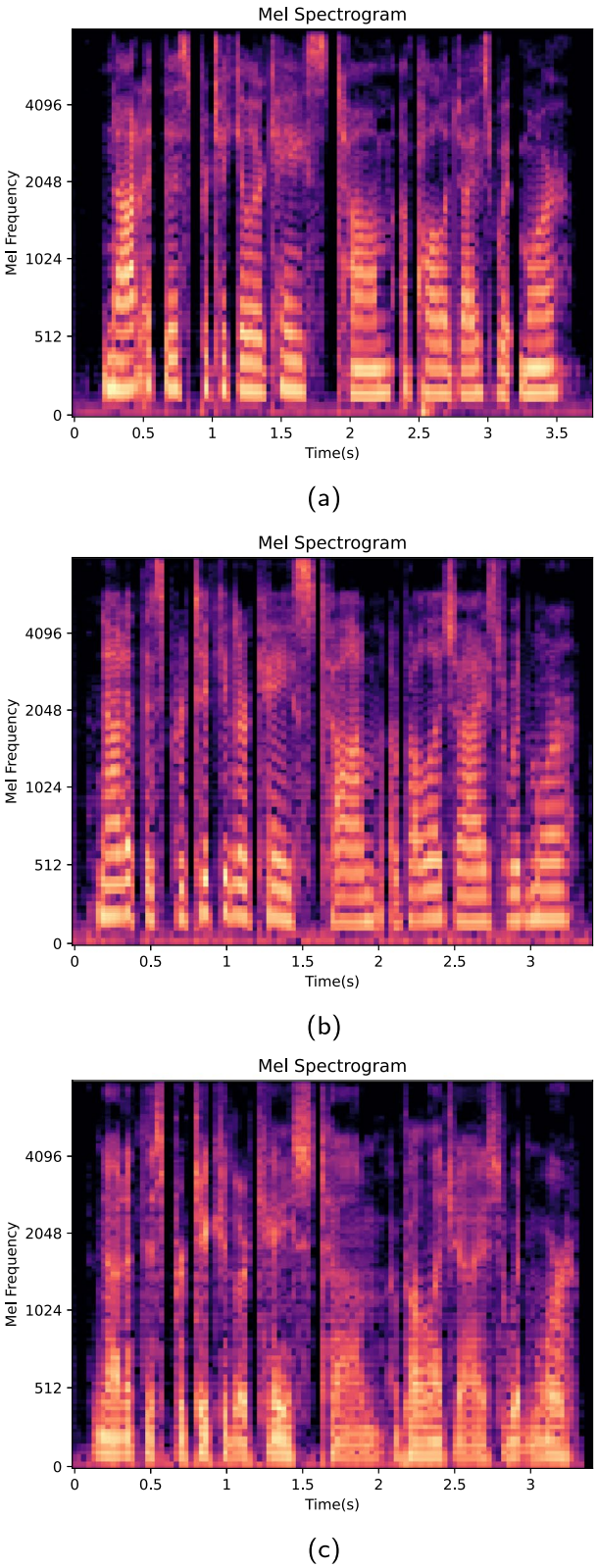

Fig. 9 **a** Mel-spectrogram of ground truth (arctic_a0002 of the source speaker "rms" in the CMU corpus). **b** W2VC cross-gender conversion, Mel-spectrogram of the target speaker "slt". **c** W2VC same-gender conversion, Mel-spectrogram of the target speaker "bdl"

similarity scores and an objective evaluation of MCD scores. Table 5 shows the influence of different modules on the MOS values of speech naturalness and speaker similarity of the network at 95% confidence intervals. As shown in Table 5, the models from both auxiliary networks outperform the models using only CTC or GRL-based speaker classification networks in terms of naturalness and similarity of intra-gender and inter-gender conversions. They are superior to a model where neither auxiliary network is used. Therefore, the presence of the auxiliary networks improves the performance of the models.

Table 6 explores the objective impact of the CTC-assisted network and the GRL-based speaker classifier network on W2VC. We used the MCD score for objective evaluation. As can be seen from Table 6, the proposed model with both auxiliary networks has the smallest MCD value and is better than the model using only CTC or GRL-based speaker classification network. Models that use neither CTC nor GRL perform the worst. The above two sets of experiments demonstrate the effectiveness of the two auxiliary content encoders proposed by W2VC in learning a purer content representation. Someone may argue the model's vulnerability to different types of target reference, e.g., speech with a high level of expressivity, or speech with noisy conditions. We think the current model might to some extent vary with these conditions. Generally, we may need to resort to additional model or method to extract diverse fine-grained information of the target speech, which could be further fully investigated in the future work.

3.9 Speech recognition results

We objectively measure the effect of CTC-assisted network module on text supervision. Speech intelligibility was measured using word error rate (WER) and character error rate (CER) to observe the regression with respect to the source audio, and the results are shown in Table 7.

As can be seen in Table 7, W2VC shows the highest recognition accuracy, CTC-VQ-VAE performs worse than W2VC, and FragmentVC achieves the worst. The experimental results show that CTC can be used as an auxiliary network to accelerate the convergence of the

Table 5 The MOS with 95% confidence intervals shows the impact of different modules in network on naturalness and similarity

Method	Nat.			Sim.		
	Intra	Inter	Avg	Intra	Inter	Avg
W2VC	4.42 ± 0.1737	4.48 ± 0.1632	4.45 ± 0.1970	3.46 ± 0.2571	3.78 ± 0.2859	3.62 ± 0.2153
w/o CTC	4.32 ± 0.1925	4.27 ± 0.2437	4.30 ± 0.2512	3.43 ± 0.3270	3.62 ± 0.3727	3.53 ± 0.1470
w/o GRL	4.22 ± 0.2431	4.32 ± 0.1930	4.27 ± 0.2281	3.41 ± 0.2372	3.77 ± 0.2202	3.59 ± 0.2881
w/o CTC+GRL	3.45 ± 0.2856	3.48 ± 0.2849	3.47 ± 0.2763	2.96 ±0.3016	2.96 ± 0.2962	2.96 ± 0.2978

Table 6 Ablation study on the CTC and GRL-based speaker classification network

Method	MCD (dB)
W2VC	8.901
w/o CTC	9.287
w/o GRL	9.328
w/o CTC+GRL	9.505

Table 7 Speech recognition accuracy of the converted speech using different VC methods in terms of WER and CER

Method	WER (↓)	CER (↓)
VQ-VAE	20.21	10.30
CTC-VQ-VAE	2.99	1.08
FragmentVC	72.85	46.10
W2VC	1.63	0.57
Ground truth	1.30	0.48

model. Moreover, the additional text supervision can jointly optimize the content encoder to learn pure content representation and improve the sound quality of the converted speech.

4 Conclusions

In this paper, we propose W2VC, a one-shot VC method based on WLMR. The traditional Mel-spectrum is used as the input content feature, and the synthesized speech quality is not high. We use WLMR and the reconstructed representation as the embedding of content information. Compared with the baseline, using WLMR reduces the loss of content information in the embedding vector and improves the speech quality. To address the impact of the content vector, we employ a supervised model based on connection timing classification and a speaker classifier based on gradient inversion layer to remove the adulterated speaker information from the content representation. From the subjective and objective tests of the experiment, these two auxiliary network modules, the purification of the content vector is complementary in performance. To address the impact of inconsistent input features of the vocoder on speech quality, we adopt the WLMR-based HiFi-GAN vocoder. From the subjective and objective evaluation of the vocoder experiments, it can be seen that using the retrained vocoder can synthesize well speech from the reconstructed WLMR, greatly improving the speech quality.

Acknowledgements
Not applicable.

Authors' contributions
Huang Hao, Lin Wang, and Jichen Yang proposed different ideas and jointly designed the model. Lin Wang did the experiments and wrote the manuscript. Hao Huang and Jichen Yang supervised the study and refined the manuscript. All authors read and approved the final manuscript.

Authors' information
Hao Huang received the B.E. degree from Shanghai Jiao Tong University, Shanghai, China, in 1999, the M.E. degree from Xinjiang University, Urumqi, China, 2004, and the Ph.D. degree from Shanghai. Jiao Tong University, in 2008, respectively. He is currently a Professor with School of Information Science and Engineering, Xinjiang University. His research interests include speech and language processing, and multi-media human-computer interaction.
Wang Lin was born in Weihai City, Shandong Province, China in 1998. She graduated from Xinjiang University with a bachelor's degree in Computer Science and technology. Enter Xinjiang University in 2020. Currently, she is a master's student in the School of Information Science and Engineering, Xinjiang University. Her current research interests include voice conversion and text-to-speech speech synthesis.
Jichen Yang received Ph.D degree in Communication and Information System from South China University of Technology (SCUT), China in 2010. From 2011 to 2015, he was a Post doctor in SCUT. From 2016 to 2020, he was a Researcher Fellow initially at the Department of Human Language Technology, Institute for Infocomm Research (I2R), A*STAR, Singapore and then in the Human Language Technology Lab, Department of Electrical and Computer Engineering, National University of Singapore, Singapore. Now he is an associate professor with the School of Cyber Security, Guangdong Polytechnic Normal University, Guangzhou and School of Electronics and Information Engineering, South China Normal University, Foshan, China. His research interests mainly include voice conversion, speech synthesis, spoofing attack detection and cross-media signal processing.
Ying Hu received the B.S. degree and M.S. degree in Electronics and Information Engineering from Xinjiang University, Urumqi, China, in 1997 and 2002, and Ph.D. degree from the School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, China, in 2016. She is currently an Associate Professor with Xinjiang University. Her current research interests include content-based audio analysis and music information retrieval.
Liang He received the B.S. degree in communication engineering from the Civil Aviation University of China, Tianjin, China, in 2004, the M.S. degree in information science and electronic engineering from Zhejiang University, Hangzhou, China, in 2006, and the Ph.D. degree in information and communication engineering from Tsinghua University, Beijing, China, in 2011. He is currently an Associate Professor with the Department of Electronic Engineering, Tsinghua University. His research interests include speech signal processing and artificial intelligence, primarily in speaker recognition, language recognition, and acoustic event detection.

Funding
This work was supported by the Opening Project of Key Laboratory of Xinjiang, China (2020D04047), Natural Science Foundation of China (61663044), the National Key R&D Program of China (2020AAA0107902).

Availability of data and materials
All datasets used in this paper are publicly available and include the VCTK [32] and the CMU ARCTIC [33].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 3 March 2023 Accepted: 12 October 2023
Published online: 28 October 2023

References

1. A. Kain, M.W. Macon, in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*. Spectral voice conversion for text-to-speech synthesis, vol. 1 (IEEE, 1998), pp. 285–288
2. K. Kobayashi, T. Toda, T. Nakano, M. Goto, G. Neubig, S. Sakti, S. Nakamura, in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Regression approaches to perceptual age control in singing voice conversion (IEEE, 2014), pp. 7904–7908
3. Z. Du, B. Sisman, K. Zhou, H. Li, in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. Expressive voice conversion: A joint framework for speaker identity and emotional style transfer (IEEE, 2021), pp. 594–601
4. C.C. Hsu, H.T. Hwang, Y.C. Wu, Y. Tsao, H.M. Wang, in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. Voice conversion from non-parallel corpora using variational auto-encoder (IEEE, 2016), pp. 1–6
5. X. Tian, J. Wang, H. Xu, E.S. Chng, H. Li, in *Odyssey*. Average modeling approach to voice conversion with non-parallel data, vol. 2018 (2018), pp. 227–232
6. T. Toda, A.W. Black, K. Tokuda, Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. IEEE Trans. Audio Speech. Lang. Process. **15**(8), 2222–2235 (2007)
7. D.Y. Wu, Y.H. Chen, H.Y. Lee, Vqvc+: One-shot voice conversion by vector quantization and u-net architecture. arXiv preprint arXiv:2006.04154 (2020)
8. D.P. Kingma, M. Welling, Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
9. L. Wan, Q. Wang, A. Papir, I.L. Moreno, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Generalized end-to-end loss for speaker verification (IEEE, 2018), pp. 4879–4883
10. L. Sun, K. Li, H. Wang, S. Kang, H. Meng, in *2016 IEEE International Conference on Multimedia and Expo (ICME)*. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training (IEEE, 2016), pp. 1–6
11. W.C. Huang, S.W. Yang, T. Hayashi, H.Y. Lee, S. Watanabe, T. Toda, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. S3prl-vc: Open-source voice conversion framework with self-supervised speech representations (IEEE, 2022), pp. 6552–6556
12. S.W. Yang, P.H. Chi, Y.S. Chuang, C.I.J. Lai, K. Lakhotia, Y.Y. Lin, A.T. Liu, J. Shi, X. Chang, G.T. Lin, et al., Superb: Speech processing universal performance benchmark. arXiv preprint arXiv:2105.01051 (2021)
13. Y.Y. Lin, C.M. Chien, J.H. Lin, H.y. Lee, L.s. Lee, in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Fragmentvc: Any-to-any voice conversion by end-to-end extracting and fusing fine-grained voice fragments with attention (IEEE, 2021), pp. 5939–5943
14. J.H. Lin, Y.Y. Lin, C.M. Chien, H.Y. Lee, S2vc: A framework for any-to-any voice conversion with self-supervised pretrained representations. arXiv preprint arXiv:2104.02901 (2021)
15. B. van Niekerk, L. Nortje, H. Kamper, Vector-quantized neural networks for acoustic unit discovery in the zerospeech 2020 challenge. arXiv preprint arXiv preprint arXiv:2005.09409 (2020)
16. A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499 (2016)
17. K. Kumar, R. Kumar, T. De Boissiere, L. Gestin, W.Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, A.C. Courville, Melgan: Generative adversarial networks for conditional waveform synthesis. Adv. Neural Inf. Process. Syst. **32** (2019)
18. J. Kong, J. Kim, J. Bae, Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. Adv. Neural Inf. Process. Syst. **33**, 17022–17033 (2020)
19. A. Graves, S. Fernández, F. Gomez, J. Schmidhuber, in *Proceedings of the 23rd international conference on Machine learning*. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks (2006), pp. 369–376
20. Y. Ganin, V. Lempitsky, in *International conference on machine learning*. Unsupervised domain adaptation by backpropagation (PMLR, 2015), pp. 1180–1189
21. X. Zhao, F. Liu, C. Song, Z. Wu, S. Kang, D. Tuo, H. Meng, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Disentangling content and fine-grained prosody information via hybrid asr bottleneck features for voice conversion (IEEE, 2022), pp. 7022–7026
22. S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao et al., Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE J. Sel. Top. Signal Process. **16**(6), 1505–1518 (2022)
23. W.N. Hsu, B. Bolte, Y.H.H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 3451–3460 (2021)
24. D. Ulyanov, A. Vedaldi, V. Lempitsky, Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
25. X. Huang, S. Belongie, in *Proceedings of the IEEE international conference on computer vision*. Arbitrary style transfer in real-time with adaptive instance normalization (2017), pp. 1501–1510
26. J.C. Chou, C.C. Yeh, H.Y. Lee, One-shot voice conversion by separating speaker and content representations with instance normalization. arXiv preprint arXiv:1904.05742 (2019).
27. J. Wang, J. Li, X. Zhao, Z. Wu, S. Kang, H. Meng, Adversarially learning disentangled speech representations for robust multi-factor voice conversion. arXiv preprint (2021)
28. A.T. Liu, S.W. Yang, P.H. Chi, P.C. Hsu, H.Y. Lee, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders (IEEE, 2020), pp. 6419–6423
29. Y. Wang, D. Stanton, Y. Zhang, R.S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, R.A. Saurous, in *International Conference on Machine Learning*. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis (PMLR, 2018), pp. 5180–5189
30. P. Wu, Z. Ling, L. Liu, Y. Jiang, H. Wu, L. Dai, in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. End-to-end emotional speech synthesis using style tokens and semi-supervised training (IEEE, 2019), pp. 623–627
31. R. Valle, J. Li, R. Prenger, B. Catanzaro, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens (IEEE, 2020), pp. 6189–6193
32. C. Veaux, J. Yamagishi, K. MacDonald. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit[J]. University of Edinburgh. The Centre for Speech Technology Research (CSTR). 6, 15 (2017).

33. J. Kominek, A.W. Black, in *Fifth ISCA workshop on speech synthesis*. The cmu arctic speech databases (2004)
34. X. Kang, H. Huang, Y. Hu, Z. Huang, Connectionist temporal classification loss for vector quantized variational autoencoder in zero-shot voice conversion. Digit. Signal Process. **116**, 103110 (2021)
35. A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations. Adv. Neural Inf. Process. Syst. **33**, 12449–12460 (2020)

Publisher's Note